Research

Open Access

Phylogenetically and spatially conserved word pairs associated with gene-expression changes in yeasts

Derek Y Chiang[*], Alan M Moses[†], Manolis Kellis[‡], Eric S Lander[§] and Michael B Eisen[¶¥]

Addresses: [*]Department of Molecular and Cell Biology, University of California, Berkeley, CA 94720, USA. [†]Graduate Group in Biophysics, University of California, Berkeley, CA 94720, USA. [‡]Whitehead/MIT Center for Genome Research, Department of Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [§]Whitehead/MIT Center for Genome Research, Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [¶]Department of Genome Sciences, Life Sciences Division, Ernest Orlando Lawrence Berkeley National Lab, 1 Cyclotron Road, Berkeley, CA 94720, USA. [¥]Center for Integrative Genomics and Division of Genetics and Development, Department of Molecular and Cell Biology, University of California, Berkeley, CA 94720, USA.

Correspondence: Michael B Eisen. E-mail: mbeisen@lbl.gov

Published: 26 June 2003

Genome ***Biology*** 2003, **4:**R43

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2003/4/7/R43

Received: 28 January 2003
Revised: 28 April 2003
Accepted: 15 May 2003

Abstract

Background: Transcriptional regulation in eukaryotes often involves multiple transcription factors binding to the same transcription control region, and to understand the regulatory content of eukaryotic genomes it is necessary to consider the co-occurrence and spatial relationships of individual binding sites. The determination of conserved sequences (often known as phylogenetic footprinting) has identified individual transcription factor binding sites. We extend this concept of functional conservation to higher-order features of transcription control regions.

Results: We used the genome sequences of four yeast species of the genus *Saccharomyces* to identify sequences potentially involved in multifactorial control of gene expression. We found 989 potential regulatory 'templates': pairs of hexameric sequences that are jointly conserved in transcription regulatory regions and also exhibit non-random relative spacing. Many of the individual sequences in these templates correspond to known transcription factor binding sites, and the sets of genes containing a particular template in their transcription control regions tend to be differentially expressed in conditions where the corresponding transcription factors are known to be active. The incorporation of word pairs to define sequence features yields more specific predictions of average expression profiles and more informative regression models for genome-wide expression data than considering sequence conservation alone.

Conclusions: The incorporation of both joint conservation and spacing constraints of sequence pairs predicts groups of target genes that are specific for common patterns of gene expression. Our work suggests that positional information, especially the relative spacing between transcription factor binding sites, may represent a common organizing principle of transcription control regions.

Background

All organisms have evolved intricate signaling networks that sense and respond to their environment. At a cellular level, the activation of one or more signaling networks often leads to coordinated changes in gene expression, via the regulated activity and binding of transcription factors to transcription control regions (TCRs) of genes (for example, enhancers and upstream activating sequences). In yeast and most other eukaryotes, the transcriptional regulation of individual genes is often multifactorial, as multiple transcription factors may bind to a single TCR [1,2]. Multifactorial regulation encompasses several distinct biochemical mechanisms. In some cases, transcription factors may bind cooperatively to adjacent DNA sites via direct physical interaction [3,4]. In other cases, multiple transcription factors that bind independently may recruit a common co-activator [5], or may act independently of one another to alter gene expression in response to distinct cellular cues [6]. Recent studies have also suggested that nearby transcription factors may collaboratively compete with nucleosomes, thus enhancing the binding of individual transcription factors [7,8]. Many experiments in yeast have shown that specific pairs of factors must be bound near each other for multifactorial regulation to occur [7–10], and it is on these spatial constraints that we focus here.

The challenges in understanding how regulatory information is encoded in genomes include both the identification of regulatory sequences in TCRs and the elucidation of sequence constraints on productive multifactorial regulation. Previous computational work has been devoted to identifying putative binding sites for transcription factors. A plethora of computational methods has been developed to find over-represented sequences in a subset of genes believed to contain a common transcription factor binding site (reviewed in [11]). The rapid pace of genome sequencing has enabled a complementary approach - phylogenetic footprinting (reviewed in [12,13]) - which recognizes that the conservation of sequences across related organisms often reflects evolutionary selection for their presence in TCRs. Several algorithms have been developed to perform systematic phylogenetic footprinting analyses [14–16].

After compiling a collection of putative binding sites, associations can be made between various binding site assortments and gene expression. Some recent approaches include Boolean logic [17], regression methods [18–21], spatial clustering [22], and multiple binding site matrix classifiers [23–25]. Spatial information on the relative locations of binding sites is ignored in all but the last two classes of approaches.

The primary aim of this work was to incorporate positional information and phylogenetic footprinting in methods of identifying sequence motifs that may regulate gene expression. Consequently, we expanded the focus of phylogenetic footprinting from the conservation of contiguous sequences to higher-order features of TCRs, namely the spatial organization of individual binding sites. Because transcription factors participating in multifactorial regulation may require binding sites in physical proximity to each other, we searched for groups of conserved sequences that were more closely spaced in TCRs than expected. We refer to these spatially organized sequences as conserved 'word templates'. As a proof of principle, we started with the simplest example of such templates: pairs of conserved words of 6 base-pairs (bp). Conservation was assessed using the genome sequences of *Saccharomyces cerevisiae* and three closely related *Saccharomyces* species, which had been sequenced to identify conserved regulatory sequences [26]. To exploit this comparative genomic data, we have devised a method that systematically tested sequence pairs for joint conservation across genomes and close spacing within individual TCRs. As genes regulated by the same set of transcription factors often display similar gene-expression patterns in certain experimental conditions, we identified pairs of conserved word templates whose gene targets were associated with common changes in gene expression. We adopted a group-by-sequence approach to first identify genes that contained the word-pair templates and then to test for significant associations with expression levels of the identified genes [27]. Significant associations between conserved word-pair templates and specific gene-expression changes and the prevalence of known transcription factor binding sites suggest that conserved word-pair templates comprise sequences important for multifactorial regulation in yeast. In addition, conserved word-pair templates represent more specific predictors of gene expression than individual words or word pairs in *S. cerevisiae*.

Results

Identification of conserved word-pair templates

Multiple genome sequences provide additional power to studies of gene regulation. Because of natural selection, mutations accumulate more rapidly in non-functional DNA regions than in functionally constrained bases. Given a multiple sequence alignment of orthologous sequences from closely related species, the aligned and invariant regions should be enriched for functionally important residues [12,13]. Additional *Saccharomyces* genomes were sequenced to ensure sufficient sequence similarity to *S. cerevisiae* such that orthologous regions could be reliably aligned, yet enough sequence divergence that functional sequences would be much more conserved than nonfunctional sequences [26,28]. To confirm that regulatory sequences were found in conserved regions, we tested a database of 47 known, nonredundant regulatory motifs and found that 35 show conservation ratios that were more than three standard deviations above that expected by chance [26,29].

We present a method to find conserved higher-order sequence templates from related *Saccharomyces* genomes (Figure 1). Our method incorporates sequential statistical tests, with each step focusing on a distinct property of

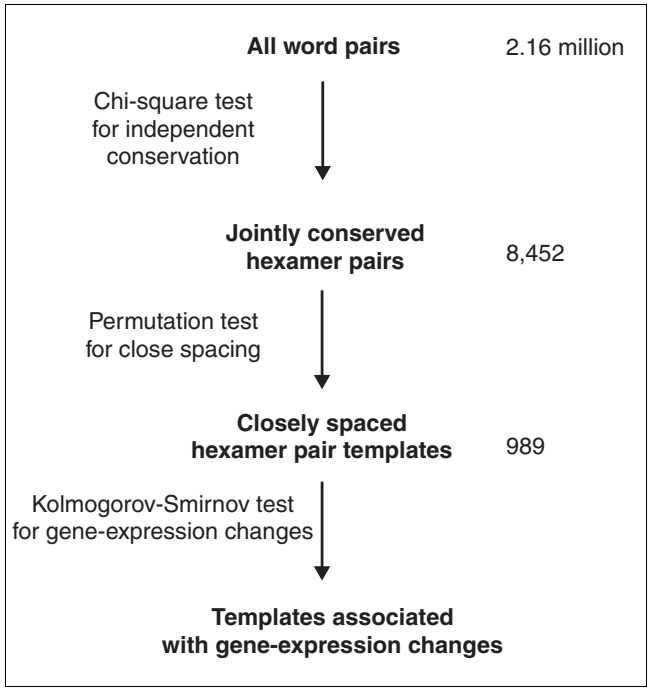

Figure 1
Overview of the method used to discover conserved word-pair templates in yeast. For the templates associated with gene-expression changes see Figure 3.

conserved sequence templates. The simplest instances of sequence templates involve word pairs and their relative spacing. As described in detail below, pairs of words that were conserved in the same intergenic regions of four *Saccharomyces* genomes were identified using a chi-square test for independence. Next, a permutation test was used to select word pairs whose physical proximity was closer than that expected by chance. Finally, to evaluate the transcriptional information contained in conserved word pairs with close spacing, the expression of genes containing TCR templates was compared to the rest of the genome. We initialized our word list using all 2,080 words of length 6, treating a given word and its reverse complement as identical. For each TCR (consisting of up to 600 bp upstream of an open reading frame), a word was labeled conserved if all six bases were identical in at least three of the four *Saccharomyces* genomes, on the basis of the CLUSTALW alignment of that TCR.

To test systematically whether words were conserved more often in the same intergenic regions of the *Saccharomyces* genomes than expected by independent conservation, a chi-square test was performed on all possible pairwise combinations of words (see Materials and methods). Pairs of words that overlapped each other by more than three nucleotides were excluded. A significant proportion of word pairs showed dependent conservation: among the 2.16 million word pairs tested, 8,452 of them (approximately 0.4%) had conservation χ^2 scores greater than 31.1. This threshold corresponds to a probability of 0.05 for obtaining one or more false positives after a Bonferroni correction for multiple testing.

Next, we selected word pairs that displayed closer physical spacing in intergenic regions than expected by chance. The choice of a statistical test to evaluate close distances must consider the local fluctuations of A+T nucleotide content in genome sequences. Previous work used a Poisson distribution to evaluate proximity between binding sites [22]. However, variability in base composition can skew occurrences of arbitrary sequences away from their expected distributions. Indeed, this statistical test was confounded by large fluctuations in the Poisson parameter estimates, which varied up to twofold within a single chromosome [22].

The effects of base-composition fluctuations, as well as varying lengths of TCRs, motivated our nonparametric statistical test for close spacing. We used the median, denoted by $\overline{D}$, to summarize a distribution of minimum distances between two words in *S. cerevisiae*. This distribution was calculated on the basis of the genes whose TCRs conserved both words, and is independent of the relative word order. If two nonoverlapping words were closely spaced in all TCRs, we should find $\overline{D}$ to be smaller than expected by chance. The statistical significance of this spacing was assessed using a permutation test by selecting the set of genes that contained a conserved word pair and then randomizing the assignment of one of the words to the genes containing that word (see Materials and methods). By permuting the TCR labels for one of the words, but not the word positions themselves, we retained the positional biases of individual words within intergenic regions.

After correcting for multiple testing by controlling the false discovery rate (FDR), a total of 989 out of 8,452 word pairs (around 12%) had significantly small values (FDR $q < 0.05$) for $\overline{D}$ (Figure 2). For a list of these closely spaced and jointly conserved word pairs see Additional data files. As a negative control, we also assayed a sample of word pairs that did not show dependent conservation (conservation $\chi^2 < 1$), yet were jointly conserved in at least 10 TCRs. No word pairs in a random sample of 42,718 pairs with nondependent conservation ($\chi^2 < 1$) showed significantly small values for $\overline{D}$, after correction for multiple testing. Figure 2 illustrates the distributions of $\overline{D}$ for conserved word-pair templates, jointly conserved word pairs, and randomly conserved word pairs. The medians of these distance distributions were 54 nucleotides, 73 nucleotides and 89 nucleotides, respectively. Notably, the median $\overline{D}$ for template pairs was significantly smaller ($p < 0.05$) than the median $\overline{D}$ for randomly conserved pairs. These results indicate that many of the word pairs that were conserved in the same intergenic regions of multiple *Saccharomyces* genomes also exhibited closer spacing in TCRs.

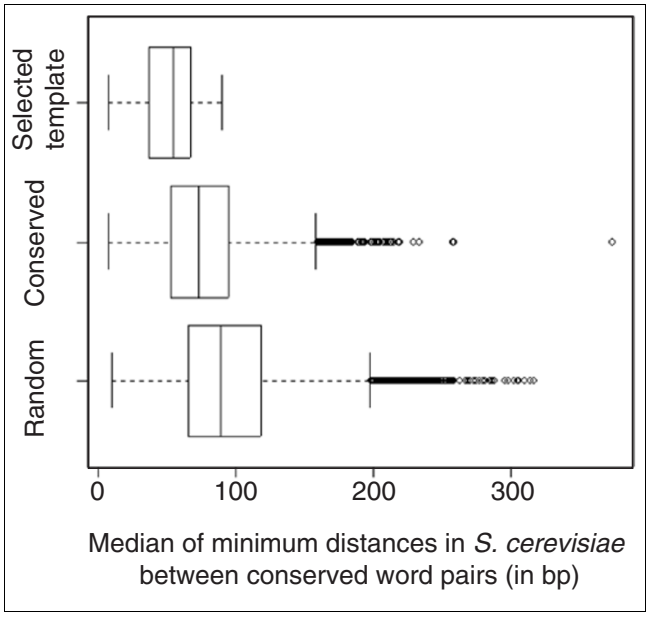

Figure 2
Word pairs in conserved word-pair templates are closely spaced in *S. cerevisiae* A comparison of the median of minimum distances is shown for $\overline{D}$ three categories of word pairs. For each category, the distribution of median of minimum distances is represented by a box-and-whisker plot, which was generated using the statistical software package R [51]; the box extends from the 25th percentile to the 75th percentile, and the vertical line within the box denotes the median of the distribution. Dashed lines extend for 1.5 times the range of the box, and circles indicate extreme values. 'Selected template' denotes closely spaced and jointly conserved word pairs ($\chi^2 > 31.1$, spacing $q < 0.05$, $N = 989$). 'Conserved' denotes dependently conserved word pairs that occur in at least 10 intergenic regions ($\chi^2 > 31.1$, $N = 3,726$) and includes all of the word pairs in the 'selected template' category. 'Random' denotes a sample of randomly conserved word pairs that occur in at least 10 intergenic regions ($\chi^2 < 1$, $N = 42,718$).

Conserved word-pair templates were significantly associated with gene expression

Our method identified conserved word-pair templates that were statistically significant with respect to both co-conservation in multiple genomes and close spacing in *S. cerevisiae* TCRs. To evaluate the regulatory information in these templates, we assessed the statistical association between gene groups that shared a template and changes in gene expression. Similarly to other group-by-sequence approaches for finding regulatory sequences, we expect that gene subsets defined by common TCR sequence features should have gene-expression patterns that are similar under conditions where the transcription factors are active, yet are different from the average expression of genes in the genome [27].

To assess the association between conserved word-pair templates and differentially expressed genes, we identified gene subsets whose TCRs contained both conserved words in the template and observed their expression patterns in *S. cerevisiae* in publicly available datasets ([30–35], see Materials and methods for details). We then conducted Kolmogorov-Smirnov (K-S) tests to evaluate for differential gene expression between each gene subset and the whole genome. K-S tests provide a nonparametric, sensitive and robust way to compare two distributions. Similar results were obtained using other statistical tests, such as *t*-tests and likelihood ratio tests (A.M. Moses, unpublished data). A $P \times C$ matrix was computed: each conserved word pair in P was assigned a K-S p-value for each experimental condition observed in C (see Materials and methods). Entries in this matrix (K-S p-values) were filtered out if the K-S p-value: did not meet the threshold for multiple testing or was less than 10 times more significant than the K-S p-value for a gene subset associated with either word alone (see Materials and methods). The latter criterion minimizes gene-expression changes that can be explained by the presence of a single conserved word.

Figure 3a shows the number of conserved word-pair templates that were significantly associated with gene-expression changes, for varying significance levels of the K-S test, which have been corrected for multiple testing (see Materials and methods). Each line indicates the number of gene subsets that were significant in a different minimum number of experimental conditions. Several hundred closely spaced word pairs were significantly associated with differential gene expression. For example, 314 word pairs met an FDR-corrected significance threshold of $p < 10^{-3}$ for five or more experimental conditions, which represented approximately 32% of all closely spaced word pairs.

The proportion of conserved word-pair templates showing significant associations with gene expression was compared to two sets of negative controls, comprising word pairs that failed either the first (co-conservation) or second (close spacing) statistical test. For the first control, we used a sample of 624 word pairs that failed the joint conservation test (conservation $\chi^2 < 1$) found in at least 25 TCRs, but also showed modest constraints on word pair spacing ($p < 0.15$). Only eight of these word pairs (approximately 1.3%) were significantly associated with gene-expression changes at an FDR-corrected threshold of $p < 10^{-3}$ for five or more experimental conditions. To assess the relative enrichment for significant associations with gene-expression changes at a variety of multiple testing thresholds, we computed an odds ratio: the proportion of significant associations among the template pairs, divided by the proportion of significant associations among the random pairs. For the above threshold, the odds ratio was about 22. In other words, gene groups that contain a common conserved word-pair template in their TCRs were about 22 times more likely to be associated with significant gene-expression changes, compared with gene groups selected using randomly conserved word pairs. As shown in Figure 3b, the odds ratios for association with gene expression changes in multiple conditions varied between 10 and 35. This analysis was repeated for a sample of 2,737 co-conserved (conservation $\chi^2 > 31.1$) word pairs that failed the close spacing test (permutation $p > 0.05$ after multiple testing), yet

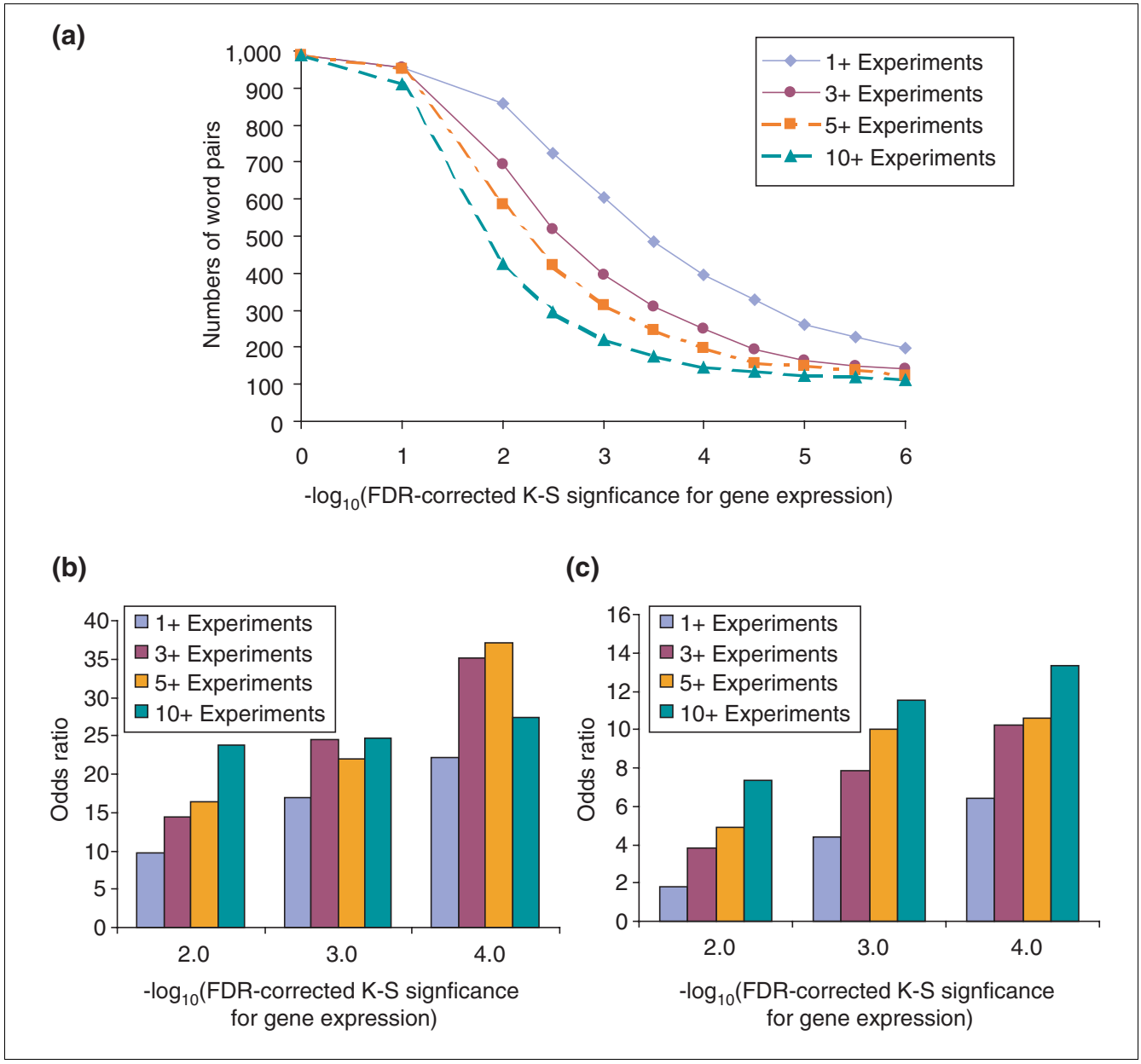

Figure 3
Conserved word-pair templates are associated with significant changes in gene expression. **(a)** Total number of conserved word-pair template associations at different K-S significance values. The horizontal axis shows different FDR-corrected significance levels for the Kolmogorov-Smirnov test (see Materials and methods). The number of closely spaced word pairs meeting this cutoff for different minimum numbers of expression conditions is shown on the vertical axis. Word pairs were also filtered for an improvement of 10× over the K-S significance from any single word. **(b)** Relative enrichment for significant gene-expression associations compared to independently conserved words. Relative enrichment was computed as an odds ratio: the fraction of gene groups selected by conserved word-pair templates associated with significant gene-expression changes, divided by the fraction of gene groups selected by randomly conserved word pairs associated with significant gene-expression changes. Templates were chosen as the set of 989 word pairs showing dependent conservation and close spacing ($\chi^2 > 31.1$, spacing $q < 0.05$); the random word pairs included 624 pairs showing independent conservation and modest spacing constraints ($\chi^2 < 31.1$, spacing $q < 0.15$). The odds ratio is shown on the vertical axis; various FDR-corrected significance thresholds for gene-expression associations are shown on the horizontal axis. Word pairs were filtered for an improvement of 10× over the K-S significance from any single word. **(c)** Relative enrichment for significant gene-expression associations compared to co-conserved words that failed the close spacing test. Templates were chosen as the set of 989 word pairs showing dependent conservation and close spacing ($\chi^2 > 31.1$, spacing $q < 0.05$); the background word pairs included 2,737 pairs showing co-conservation, but no significant close spacing constraints ($\chi^2 > 31.1$, spacing $q > 0.05$). The odds ratio is shown on the vertical axis; various FDR-corrected significance thresholds for gene-expression associations are shown on the horizontal axis. Word pairs were filtered for an improvement of 10× over the K-S significance from any single word.

occurred in at least 10 intergenic regions. The relative enrichment for gene expression associations in closely spaced words is displayed in Figure 3c. Among co-conserved word pairs, those pairs that were closely spaced than expected were still about 2 to 12 times more likely to be significantly associated with gene expression changes, compared to word pairs that were not found to have significantly close spacing. We confirmed that gene groups associated with significant gene expression changes did not have statistically significant differences in their TCR sizes, as assessed by a permutation test (data not shown). Thus, gene groups that contained co-conserved and spatially close word pairs are more significantly associated with gene-expression changes than expected by chance.

Many identified sequences represented known transcription factor binding sites

In addition to their statistical significance, many conserved word-pair templates that were most strongly associated with gene-expression changes were consistent with biological information on the transcription factors known to bind those sites [36]. In all analyses described below, we used a set of 314 word pairs that had significant associations with gene-expression changes at an FDR-corrected multiple testing threshold of $p < 10^{-3}$ for five or more experiments. For visualization purposes, we organized the $P \times C$ matrix by hierarchically clustering the K-S p-values for the 314 word pairs (see Materials and methods for details).

Hierarchical clustering of this output matrix identified groups of word pairs with similar K-S p-values in specific subsets of experimental conditions (Figure 4). For these clustered output matrices, see Additional data files. In many cases, the word pairs that clustered together also comprised overlapping hexamer sequences, suggesting that some of the hexamers in different pairs may represent a larger, somewhat variable sequence (Table 1). For example, group 9 in Figure 4 included six word pairs. In each of these word pairs, one of the component words - such as TCACGT, CACGTG, or ACGTGC - matched part of the consensus binding site for Cbf1p (TCACGTG). The other component word in each pair - such as ACTGTG, CTGTGG, TGTGGC or GTGGCT - represented part of the known Met31/32p binding site (AAACTGTGG). Therefore, genes whose TCRs contained any word pair within this group probably contained a conserved Cbf1p-binding site, along with a conserved Met31/32p-binding site, and the distances between the conserved sites in these genes were also smaller than expected by chance. These results agree with the known interaction of Cbf1p and Met31/32p for the regulation of genes involved in sulfur utilization (see Discussion).

Table 1 shows a partial list of the 14 most significant groups of consensus sequences, which were assembled by joining adjacent word pairs in the clustered output matrix with overlapping sequences. Many of these consensus sequences matched transcription factor binding sites that had been biochemically verified. Several pairs of transcription factors, denoted by stars in Table 1, were not previously known to act on the same sets of target genes.

Conditions with significant gene-expression changes coincided with transcription factor activity

Further support that templates contain transcriptional regulatory information was obtained from a key observation: the experimental conditions with significant gene-expression changes often corresponded to conditions in which the cognate transcription factors are known to be active (Table 2). In addition, many gene subsets that shared an individual word-pair template in their TCRs were highly enriched for gene-expression changes. We will survey examples of word-pair templates associated with gene-expression changes, focusing our attention on several environmental stress conditions. The environmental stress response represents a paradigm for multifactorial control of transcription regulation. Genome-wide expression studies found that around 300 genes were induced and around 600 genes were repressed in response to a wide variety of stressful environmental transitions [30,37]. Many of these genes also showed subtly different expression patterns in response to specific stimuli, suggesting that the common environmental stress response may be modulated by the activity of condition-specific transcription factors [30].

Over a third of the conserved word-pair templates were associated with gene-expression changes in multiple environmental stress conditions (Figure 4a, Tables 1, 2). The largest group of overlapping word pairs contained matches to the PAC and RRPE motifs, which were associated with genes that were repressed in multiple stresses [30,38]. These motifs were discovered by their enrichment among the approximately 600 genes that were commonly repressed in stress, yet the

Figure 4 *(see following page)*
Specific patterns of gene-expression changes are associated with templates. Conserved word-pair templates are shown with significant associations with gene-expression changes in **(a)** multiple environmental stress conditions or **(b)** distinct subsets of environmental conditions. The $P \times C$ matrix of K-S p-values was hierarchically clustered by rows and visualized with TreeView [52]. Each row corresponds to a conserved word-pair template, and each column represents a single gene-expression experiment. The experimental conditions are indicated by the color bar above each panel, according to the key. The value in each cell corresponds to the K-S p-value of gene-expression changes in each condition (column) for a group of genes that contain the conserved word-pair template (row) in their TCRs. Orange denotes a K-S p-value below the FDR critical value of 0.001 for multiple testing, while gray represents values that were not significant. Word pairs that failed to meet a FDR critical value of 0.001 for multiple testing in five or more experiments are not shown. Some of the most significant conserved word-pair associations are labeled and annotated in Tables 1 and 2.

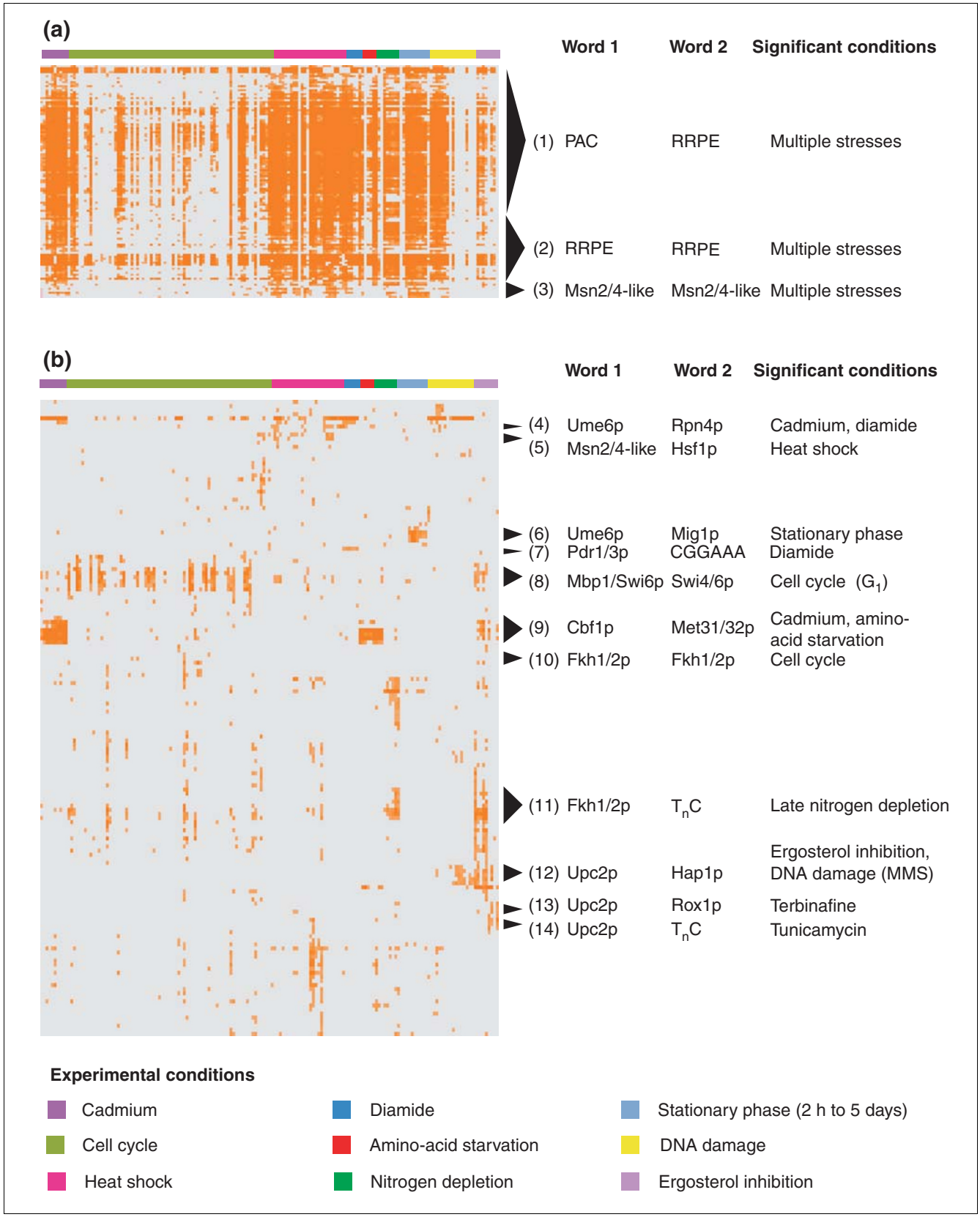

Figure 4 *(see legend on previous page)*

Table 1

Consensus sequences for the most significant groups of word pairs

	Hexamer list for word 1	Compiled sequence 1	TF for consensus 1	Hexamer list for word 2	Compiled sequence 2	TF for consensus 2	Number of word pairs
1	GAGATG GCGATG AGATGA CGATGA GATGAG ATGAGA ATGAGC TGAGAT TGAGCT GAGATG AGATGA AGCTCA	GMGATGAGMTSA	Unknown (PAC motif [38])	TGAAAA GAAAAA AAAAAT AAAATT AAATTT	TGAAAATTT	Unknown (RRPE motif [38])	75
2	AAGTGA AATGAA AGTGAA ATGAAA CTGAAA TGAAAA	ANTGAAAAA	Unknown (RRPE motif [38])	GAAAAA GAAAAT AAAATT AAATTT	GAAAAWTT	Unknown (RRPE motif [38])	40
3	GTTCCC CTCCCC ACCCCT TCCCCT	GYWCCCCT	(motif 38 [26])	CCCTTT CCTTTT CCTTAT	CCCTTWT	(motif 38 [26])	5
4*	GGCGGC GCGGCT	GGCGGCT	Ume6p	GTGGCA GGCAAA	GTGGCAAA	Rpn4p	2
5	CCCTTT CCTTTT	CCCTTTT	Msn2/4p-like	GGAGAA GGGAAA	GGRGAAA	Hsf1p	2
6	CGGCGG	CGGCGG	Ume6p	TACCCC ACCCCA CCCCAA	TACCCCAA	Mig1p	3
7*	CCGCGG	CCGCGG	Pdr1/3p	CGGAAA	CGGAAA	Unknown	1
8	AAACGC GACGCG AACGCG ACGCGT ACGCGA TCGCGT CGCGTC	ARWCGCGW	Mbp1p	CGCGAA ACGAAA GCGAAA CGAAAC CGAAAA	CRCGAAAM	Swi4/6p	9
9	TCACGT CACGTG ACGTGC	TCACGTGC	Cbf1p	ACTGTG CTGTGG TGTGGC GTGGCT	ACTGTGGCT	Met31/32p	6
10	TATTTT TTTTGT TTTGTT ATTGTT	TWTTGTT	Fkh1/2p	TGTTTA GTTTAC	TGTTTAC	Fkh1/2p	4
11	TTTGTT TTGTTT	TTTGTTT	Fkh1/2p	TTTTTC TTTTTT	TTTTTY	T_nC	4
12*	TCGTTT CGTTTA	TCGTTTA	Ecm22p | Upc2p	CCGATA CGATAA	CCGATAA	Hap1p	4
13	TCGTTT CGTTTA	TCGTTTA	Ecm22p | Upc2p	TATTGT ATTGTT	TATTGTT	Rox1p	2
14	CGTTTC GTTTCT	CGTTTCT	Ecm22p | Upc2p	TTCTTT TCTTTT CTTTTT	TTCTTTTT	T_nC	5

See legend on next page

Table 1 *(see table on previous page)*

The output ***P*** × ***C*** matrix of word pairs (***P***) that were significantly associated ($p < 0.001$) with at least five or more environmental conditions (***C***) was ordered using hierarchical clustering. Numbers correspond to groups of overlapping word pairs indicated in Figure 4. Asterisks denote sequence pairs whose involvement in multifactorial regulation has not been previously reported. Compiled sequences were assembled from groups of word pairs that were found in adjacent rows in the ordering of K-S *p*-values. As individual words must have passed all three statistical tests to be included in the output matrix, these consensus sequences may not reflect the actual biological specificities of conserved transcription factor binding sites (refer to [26,36] for a more complete list). Residues are shown in bold if they are invariant in at least two hexamers. Numbers denote the groups that are indicated in Figure 4. Multiple transcription factors that may bind the same sequence motif are separated by |. IUPAC codes used: K (G or T); M (A or C); R (A or G); S (C or G); W (A or T).

putative transcription factors that bind these sequences have yet to be determined. The second largest group of overlapping word pairs corresponded to the RRPE core, which is 10 nucleotides long, along with some flanking conserved bases. These repressed genes were enriched for rRNA processing genes, the group of genes in which this motif was originally identified [38]. Nine conserved word-pair templates contained sequences that matched most of the stress response element (STRE), the consensus site for the general stress transcription factors Msn2p/Msn4p. Genes that conserved both these words in their TCRs were significantly associated with gene-expression induction in multiple environmental stresses, including cadmium, heat shock, amino-acid starvation, nitrogen depletion and stationary phase. In most cases, the sequences comprising the word pairs were mutually overlapping. We interpret these sequences as representing different halves of the same binding site. Because our test for close spacing required non-overlapping sequences, the two words must have appeared over 6 bp away in TCRs. Thus, these genes have probably conserved at least two Msn2/4p-like consensus sequences in their TCRs.

Several groups of conserved word-pair templates only showed significant associations with gene expression changes in different subsets of stress conditions (Figure 4b, Tables 1, 2). For example, binding sites for Cbf1p and Met31/32p were found to co-occur in several conserved word-pair templates. Genes that contained conserved binding sites for these transcription factors in their TCRs were strongly induced in cadmium, amino-acid starvation and early nitrogen depletion. These conditions are consistent with the biological activity of these transcription factors, which induce transcription of sulfur-assimilation genes in response to the demand of sulfur-containing metabolites [5,39,40]. In another example, several word pairs comprising binding sites for the transcriptional repressors Mig1p and Ume6p were associated with induced gene expression in stationary phase. The observed derepression of Mig1p and Ume6p targets in stationary phase is consistent with the nuclear export of the Mig1p repressor under glucose limitation, as well as recent findings that carbon source genes can be Ume6p targets [41]. In addition, genes containing a conserved sequence similar to the consensus for Msn2/4p, an inducer of the environmental stress response, and the heat-shock transcription factor Hsf1p were significantly induced under heat shock. Once again, the conditions with most significant gene-expression changes corresponded to the known activities of the transcription factors involved.

Enrichment for known transcription factor targets among individual gene groups

Some groups of genes with shared word-pair templates were enriched for known targets of transcription factors. The vast majority of genes with conserved sites for both the Cbf1p and Met31/32p transcription factors were induced more than fourfold in cadmium, amino-acid starvation and early nitrogen depletion (Figure 5a). Half of these genes have confirmed roles in sulfur-utilization processes, such as methionine metabolism, sulfate assimilation, sulfate transport and sulfur amino acid metabolism [42]. Compared to the rest of the genome, the group of genes that conserved both of these words within their TCRs was highly enriched for sulfur-utilization genes (hypergeometric p-value $< 1 \times 10^{-16}$, after Bonferroni correction for multiple testing). In addition, we found three genes in this group (*GSH1*, *RAD59* and *BNA3*) with highly correlated expression under the above conditions, and thus may be commonly regulated by Cbf1p and Met31/32p, despite their lack of direct annotation as sulfur-utilization genes. The shared conservation of both the Cbf1p and Met31/32p sites provides further evidence that these genes comprise part of the cellular response to the demand for products of this pathway.

Genes with a conserved half-site for the Hap1p transcription factor, as well as a conserved Ecm22p/Upc2p binding site in their TCRs, were significantly associated with induction in the presence of drugs that inhibited ergosterol biosynthesis (Figure 5b). This group of 30 genes contained eight ergosterol biosynthesis genes; this proportion represented an enrichment compared to the rest of the genome (hypergeometric p-value $< 6 \times 10^{-6}$ after Bonferroni correction for multiple testing). The transcription factors Ecm22p and Upc2p have been shown to induce the expression of ergosterol biosynthesis genes in response to low intracellular concentrations of ergosterol, whereas Hap1p is known to regulate the expression of these genes according to the availability of heme and oxygen, which are required for the pathway [43,44].

Table 2

Summary for most significant groups of conserved word pairs

			Most significant word pair in consensus group			
	Conserved word pairs (compilation of overlapping words)	Known transcription factors or motifs	(χ^2, *p*-value via Bonferroni)	Median of min distance $\overline{D}$	Number of TCRs	Expression conditions with significant gene subsets (FDR significance)
1	G[AC]GATGAG, TGAAAATTT	PAC, RRPE	240.6 (10^{-49})	19±0.5	162	Repressed in multiple environmental stresses (10^{-6})
2	ANTGAAA, GAAAAWT	RRPE (Overlap)	96.9 (2×10^{-16})	43±11	68	Repressed in multiple environmental stresses (10^{-6})
3	CTCCCC, CCCTTA	Msn2/4p-like, (Overlap)	53.8 (5×10^{-7})	28±3.7	15	Induced in multiple environmental stresses (10^{-6})
4	GGCGGGC, GTGGCA	Ume6p, Rpn4p	43.7 (9×10^{-5})	48±16	25	Cadmium, diamide (10^{-4}) MMS, heat shock (10^{-3})
5	CCTTTT, GAGAAA	Msn2/4p, Hsf1p	56.2 (2×10^{-7})	54±5.4	69	Heat shock (10^{-4})
6	CCGCCG, ACCCCA	Ume6p, Mig1p	41.9 (2×10^{-4})	17±1.5	14	Stationary phase (10^{-6})
7	CCGCGG, CGGAAA	Pdr1/3p, Unknown	111 (2×10^{-19})	44±12	21	Diamide (10^{-3})
8	RACGCG, RCGAAA	Swi6p/Mbp1p, Swi4/6p,	83.0 (7×10^{-13})	33±5.0	33	Cell cycle, G1 phase (10^{-6})
9	GCACGTGC, ACTGTGGC	Cbf1p | Pho4p, Met31/32p	37.4 (2×10^{-3})	22±2.5	22	Cadmium (10^{-6})
10	T[AT]TTGTT, TGTTTA	Fkh1/2p (Overlap)	51.1 (2×10^{-6})	57±6.9	48	Cell cycle (10^{-3})
11	TTTGTT, TTTTTY	Fkh1/2p, T_nC	37.6 (2×10^{-3})	49±4.4	267	Late nitrogen depletion (10^{-3})
12	CCGATA, TCGTTT	Hap1p, Ecm22p | Upc2p	36.2 (4×10^{-3})	41±5.9	28	Ergosterol inhibition (10^{-4}) MMS (DNA damage) (10^{-3})
13	TCGTTT, TATTGTT	Rox1p, Ecm22p | Upc2p	58.8 (4×10^{-8})	55±0.5	69	Early menadione (10^{-3})
14	TGACTC, TCTTTT	Gcn4, T_nC	35.6	59±9.1	63	Amino-acid starvation (10^{-5})

Statistics are listed for one representative word pair for each group of overlapping word pairs, numbered as in Figure 4. Multiple transcription factors that may bind the same sequence motif are separated by |. To summarize the close spacing ($\overline{D}$) between conserved word pairs, we report the median of the distribution of minimum distances in *S. cerevisiae* ± standard deviation of the medians of the distribution of minimum distances in all four *Saccharomyces* genomes.

Conserved word pairs were more informative than sequence features derived from single words or single species

The above results from the K-S test strongly suggested that sequence features based on the co-conservation and close spacing of word pairs identified examples of multifactorial regulation. Two other statistical tests were also performed to examine how information contained in conserved word-pair templates compared to other sequence features derived from *S. cerevisiae*, or from single conserved words. Informative sequence features should be both highly specific (a high proportion of genes that share the feature should show gene-

Figure 5 *(see following page)*
Enrichment for known transcription factor targets among individual gene groups. Gene-expression patterns are shown for genes whose TCRs contain the known binding sites for: **(a)** Cbf1p (CACGTG) and Met31/32p (TGTGGC); or **(b)** Hap1p (CCGATA) and Ecm22p/Upc2p (TCGTTT). The genes are listed in ascending order of minimum distance between the two conserved words in the corresponding TCRs of *S. cerevisiae*. Each row in these diagrams represents a given gene's expression pattern under the conditions shown in each column: exposure to increasing concentrations of cadmium chloride (from 0.05 mM to 0.4 mM); an amino-acid starvation timecourse; a nitrogen-source depletion timecourse [30]; and growth in the presence of drugs or genetic alterations that inhibit ergosterol biosynthesis (*erg3*Δ, itraconazole, *erg28*Δ, overexpressed *ERG11*, *erg2*Δ, tunicamycin, terbinafine, *erg6*Δ, overexpressed *HMG2*) [35]. A red color indicates that the gene's expression was induced under those conditions, while a green color indicates that the gene was repressed under those conditions; black indicates no detectable change in expression, and gray indicates missing data. Gene names in purple correspond to genes with confirmed roles in (a) sulfur utilization or (b) ergosterol biosynthesis; gene names in orange show highly correlated expression patterns, despite their lack of annotation as sulfur-utilization genes. Arrows above the columns indicate conditions in which the displayed gene groups show significant gene-expression changes according to the K-S test, FDR correction for multiple testing at a *p*-value of 0.001.

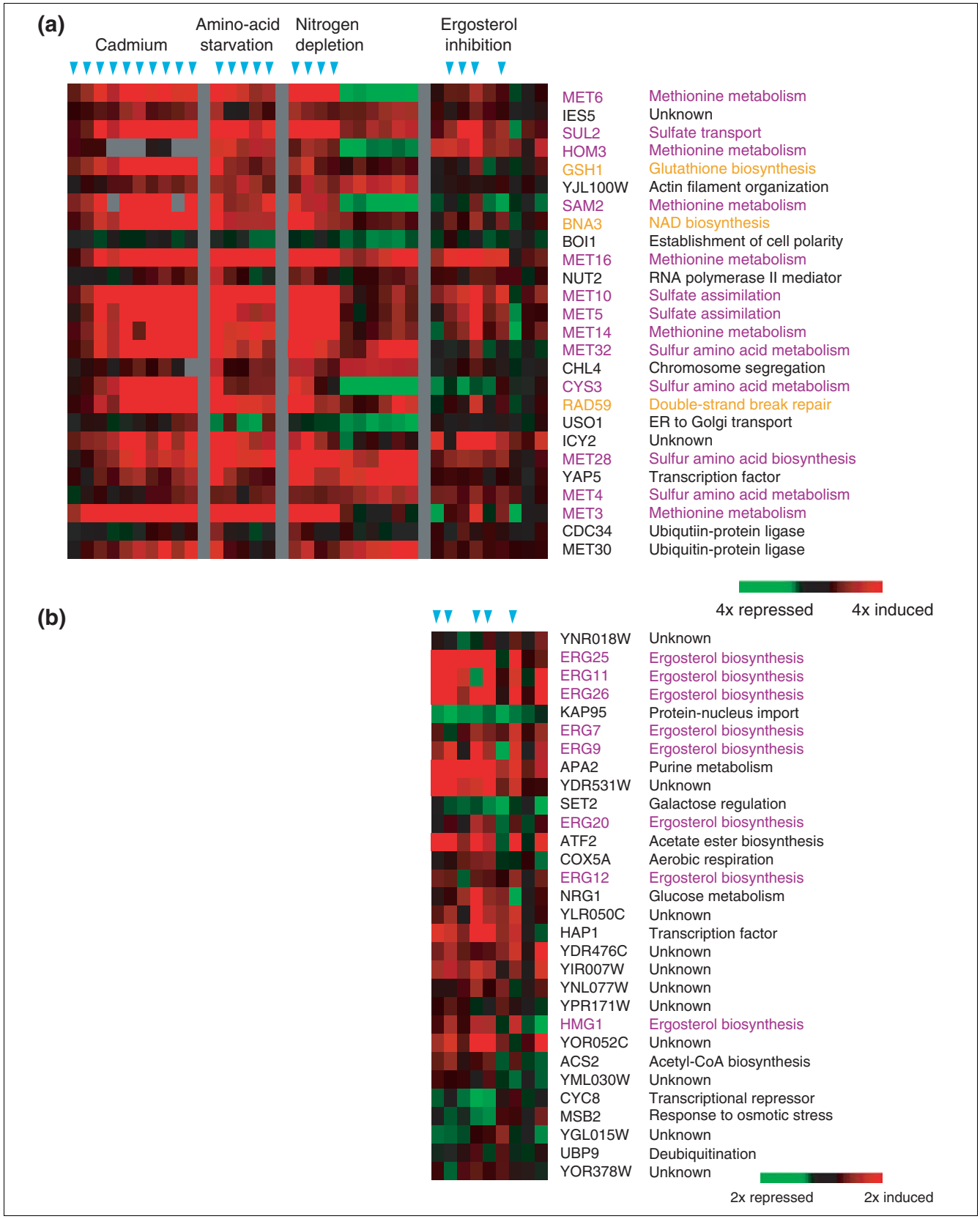

Figure 5 *(see legend on previous page)*

expression changes) and highly sensitive (most of the genes that show gene-expression changes should also share the feature).

To assess the specificity of a sequence feature in explaining gene expression, we computed the average expression profile for all genes that shared that feature. We expect that if a sequence feature represented a transcription factor binding site, genes containing that feature in their TCRs would be induced (or repressed), on average, compared to all the genes in the genome [27]. By comparing the average expression profile derived from conserved word-pair templates with average expression profiles derived from simpler sequence features, we assessed how much information was obtained by incorporating both the conservation and pairwise combination of sequences. For representative word pairs associated with significant gene-expression changes in environmental stress conditions, we compared the average expression profile for: gene subsets that shared single words in *S. cerevisiae*; gene subsets that conserved single words among multiple genomes; and gene subsets that shared both words in *S. cerevisiae* (Figure 6 and see Additional data files). In general, the average gene-expression profiles for conserved word pairs were more significantly different from the average expression of genes in the genome when either conservation or word pairs was used as an additional criterion for gene selection. In Figure 6, the last two rows for each word pair indicate the average expression profiles for genes that shared both words in *S. cerevisiae*, as well as the average expression profile for genes that conserved both words in multiple genomes, respectively. Strikingly, the consideration of word-pair conservation yielded further increases in average gene-expression profiles compared to word pairs in *S. cerevisiae* alone. Similar effects were observed for the PAC-RRPE and the overlapping RRPE pairs associated with genes repressed in the environmental stress response (see Additional data files). Thus, conserved word-pair templates contained more specific predictors of gene expression than comparable sequence templates derived from *S. cerevisiae* alone.

To evaluate how well sequence features can explain gene expression changes across the whole genome, several groups have constructed linear regression models using various choices for features [18–21]. The *R*-square statistic of a regression model indicates the percent of global variance that can be explained using the sequence features in the model. Models with better fits to the genome-wide expression data would thus have greater *R*-square values. To assess the sensitivity of individual word pairs in explaining global gene expression, we first constructed regression models using individual word pairs (see Materials and methods). We chose three representative environmental conditions: amino-acid starvation (30 min) [30]; stationary phase (10 h in YPD) [30]; and ergosterol inhibition (terbinafine) [35]. We constructed regression models using counts of individual words in *S. cerevisiae* TCRs, or using counts of words that were conserved among *Saccharomyces* TCRs. Sequence conservation improved the fit of regression models based on individual word pairs (ΔR^2 = 0.3-1.3%) (see Additional data files). These results clearly show that sequences conserved in multiple *Saccharomyces* species were more likely to be associated with gene-expression changes.

To assess the joint contribution of word pairs on gene expression, we also included interaction terms between the individual words only if their coefficients were statistically significant (see Materials and methods). Pairwise interaction terms, expressed as the product of scores for two features, represent a standard way to assess whether two features contribute nonadditively to gene expression [19]. Indeed, the inclusion of significant pairwise interaction terms improved the fits for both the *S. cerevisiae* sequence model and the conserved sequence model, increasing the *R*-square by a further 0.2% to 0.9% (see Additional data files). Whereas the interaction terms only comprise a small proportion of the global variance, they can be interpreted as statistical evidence of dependence between sequence features [19]. Therefore, the nonadditive contributions of conserved word-pair templates further suggest their involvement in multifactorial regulation.

We expanded these models to include multiple conserved word-pair templates using a stepwise linear regression procedure. The set of potential sequence features was expanded to

Figure 6 *(see following page)*
Incorporation of conservation and word pairs provided more informative average expression profiles. Groups of genes whose TCRs contained various sequence features were summarized by the average of their gene-expression profiles. Each row in these diagrams represents a given gene group's average expression pattern under the conditions shown in each column: exposure to increasing concentrations of cadmium chloride (from 0.05 mM to 0.4 mM); 20 min after heat shock to 37°C (from 17°C, 21°C, 25°C, 29°C, and 33°C); an amino-acid starvation time course; a nitrogen-source depletion time course; progression into stationary phase (2 h, 4 h, 6 h, 8 h, 10 h, 12 h, 1 day, 2 days, 3 days, 5 days of growth) [30]. Representative conserved word-pair templates were chosen for analysis, corresponding to: **(a)** Msn2/4p-like sequences (CTCCCC and CCCTTA); **(b)** Cbf1p (CACGTG) and Met31/32p (TGTGGC) binding sites; **(c)** Mig1p (ACCCCA) and Ume6 (CCGCCG) binding sites; **(d)** Msn2/4p-like (CCCCTT) and Hsf1p-like (GAGAAA) sequences. For each of the panels (a-d) each row represents the average expression profile for gene groups chosen by different sequence features in their TCRs: single words found in *S. cerevisiae* (rows 1 and 2); single words conserved in three or more *Saccharomyces* genomes (rows 3 and 4); word pairs found in *S. cerevisiae* (row 5); word pairs conserved in three or more *Saccharomyces* genomes (row 6). Arrows above the columns indicate conditions under which gene groups sharing the conserved word-pair template (row 6) were significantly associated with gene-expression changes, at a *p*-value of 0.001 (K-S test after FDR correction for multiple testing).

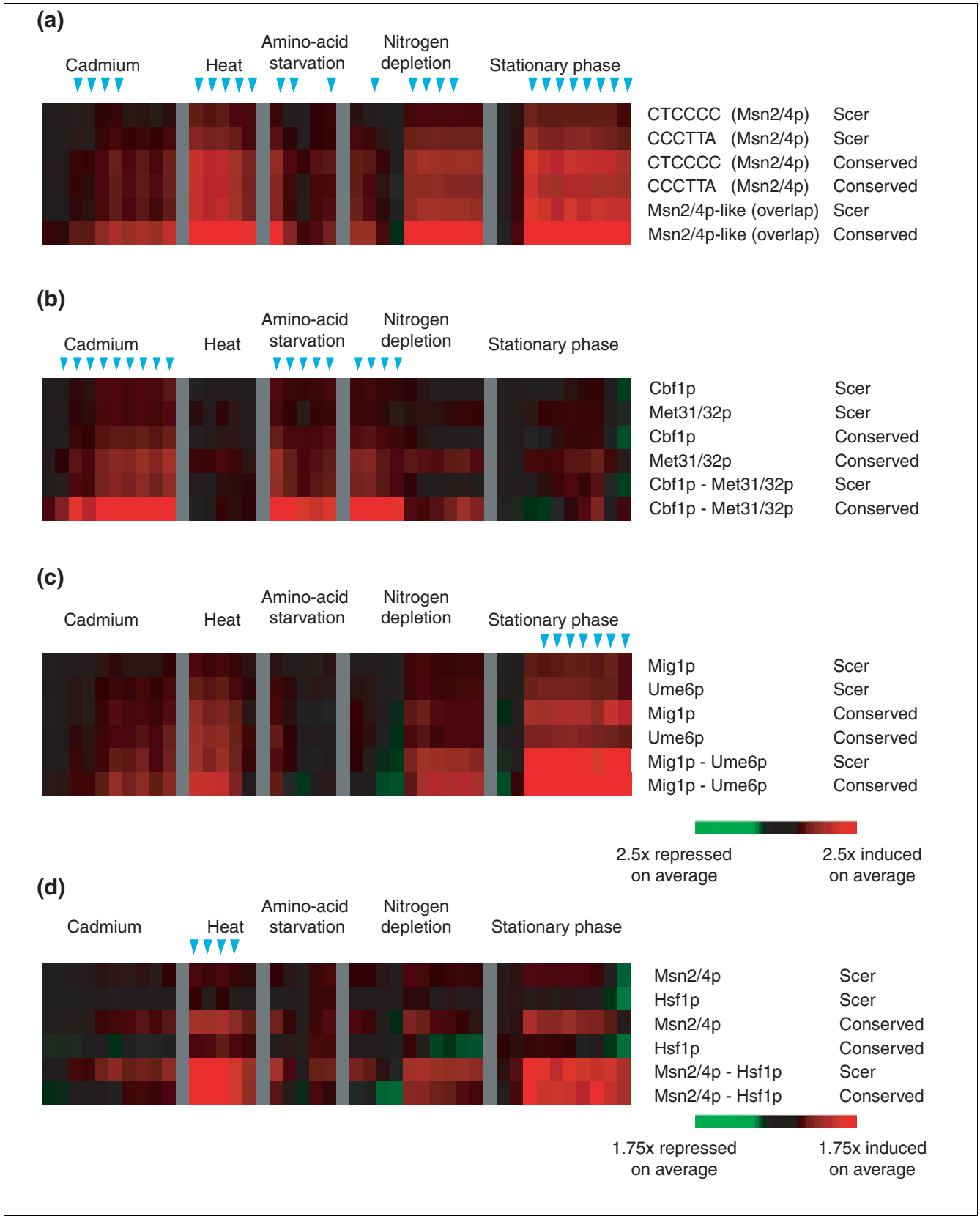

Figure 6 *(see legend on previous page)*

Table 3

Stepwise linear regression statistics

		S. cerevisiae			Three or more genomes				
	Word	M_{fc}	*p*-value	ΔR^2	M_{fc}	*p*-value	ΔR^2	*p*-value	*p*-value
Amino-acid	AAATTT	-0.165	< 2.0e-16	3.1%	-0.293	< 2.0e-16	6.6%	1.3e-37	2.0e-07
starvation	GATGAG	-	-	-	-0.333	6.7e-16	4.1%	1.6e-30	8.5e-03
0.5 h	AAGGGG	0.209	3.2e-14	1.8%	0.455	< 2.0e-16	3.5%	1.6e-22	7.9e-05
	TGTGGC	0.094	1.1e-03	0.6%	0.283	2.9e-07	1.6%	5.0e-07	8.3e-13
	CCCTTA	0.300	**2.0e-16**	**1.7%**	0.363	**< 2.0e-16**	**1.4%**	**3.2e-06**	**3.5e-03**
	TGACTC	0.229	**4.6e-10**	**0.8%**	0.311	**2.2e-11**	**1.0%**	**4.6e-01**	**1.1e-03**
	AAATTT • GATGAG	-	-	-	-0.266	**9.1e-09**	**0.5%**	-	-
	CACGTG	0.045	3.5e-01	0.5%	0.146	8.6e-03	0.5%	**1.9e-07**	**1.2e-08**
	CACGTG • TGTGGC	0.443	**3.8e-10**	**0.5%**	0.749	**1.0e-12**	**0.9%**	-	-
	GTGAAA	-0.066	**1.1e-03**	**0.3%**	-0.082	**4.6e-03**	**0.1%**	-	-
	TCTTTT	-0.022	**2.3e-02**	**0.1%**	-	-	-	-	-
	Total ΔR^2			**9.6%**			**20.2%**		
Stationary	AAATTT	-0.218	< 2.0e-16	3.2%	-0.377	< 2.0e-16	5.8%	5.5e-39	N/R
phase	AAGGGG	0.233	1.7e-11	0.9%	0.591	< 2.0e-16	4.0%	4.5e-26	N/R
YPD 10 h	**CCCTTA**	0.460	**< 2.0e-16**	**3.7%**	0.579	**< 2.0e-16**	**2.2%**	3.0e-07	N/R
	GATGAG	-	-	-	-0.242	4.1e-06	1.8%	4.4e-18	N/R
	ACCCCA	0.224	**3.0e-03**	**0.3%**	0.459	**1.5e-06**	**1.0%**	-	N/R
	AAATTT • GATGAG	-	-	-	-0.287	**1.7e-06**	**0.4%**	-	N/R
	CCGCCG	0.333	**5.1e-07**	**0.8%**	0.208	**1.5e-02**	**0.3%**	-	N/R
	ACCCCA • CCGCCG	0.294	**1.8e-02**	**0.1%**	0.807	**5.6e-05**	**0.3%**	-	N/R
	GTGAAA	-0.090	**4.2e-04**	**0.2%**	-0.122	**1.0e-03**	**0.2%**	-	N/R
	Total ΔR^2			**9.4%**			16.0%		
Terbin-	**TGACTC**	0.162	**< 2.0e-16**	**3.5%**	0.261	**< 2.0e-16**	**5.1%**	1.3e-14	N/R
afine 3 h	TCGTTT	0.071	< 2.0e-16	2.0%	0.132	< 2.0e-16	3.3%	2.5e-24	N/R
	TGAAAC	-0.055	**1.3e-12**	**1.1%**	-0.077	**9.50e-11**	**0.9%**	4.0e-03	N/R
	GATGAG	-0.029	**1.7e-03**	**0.3%**	-0.047	**6.70e-06**	**0.4%**	-	N/R
	AAGGGG	0.025	**1.1e-02**	**0.1%**	0.050	**5.40e-04**	**0.3%**	2.4e-01	N/R
	CCGATA	-0.008	**6.5e-01**	**0.1%**	0.004	**8.6e-01**	**0.1%**	-	N/R
	CCGATA • TCGTTT	0.080	**9.4e-06**	**0.3%**	0.146	**3.2e-07**	**0.5%**	-	N/R
	CCCTTA	-0.021	**5.0e-02**	**0.1%**	-0.038	**1.1e-02**	**0.1%**	-	N/R
	Total ΔR^2			**7.6%**			10.8%		

Words and pairwise interaction terms are reported in the order of selection by the stepwise linear regression procedure performed on conserved words. The influence terms (M_f), associated *p*-values, and increase in *R*-square values were computed using the statistical package R [51]. Wang *et al.* [20] and Conlon *et al.* [21] previously fit regression models using sequence features derived from *S. cerevisiae*. The *p*-values of the most similar sequences features in their regression models were reported where available; sequence features that were more significant in this analysis are indicated in bold. Dashes indicate sequence features that were insignificant in the Wang *et al.* [20] or Conlon *et al.* [21] analyses. 'N/R' indicates gene-expression data that were not analyzed by Conlon *et al.* [21]

include all words found in templates associated with significant gene expression changes in that condition, as assessed previously by the K-S test (see Materials and methods). The final *R*-square values for regression models based on occurrences of multiple words in *S. cerevisiae* ranged from 7.2% to 9.3% (Table 3 and see Additional data files). Once again, the use of conserved instances of individual words yielded better model fits, with improvements in *R*-square values from 3.1% to 9.5%. Further improvements in the model fit (ΔR^2 = 0.5% to 1.4%) were obtained using pairwise interaction terms between individual words found in the same word-pair template. The total *R*-square values for the regression models

based on conserved word-pair templates with interaction terms thus ranged from 10.8% to 20.2% (Table 3). Thus, sequence features based on conserved word-pair templates could explain more of the global gene expression changes than features based on *S. cerevisiae* alone.

Discussion

This work describes two principles for analyzing combinations of regulatory sequences. First, sequence conservation among closely related yeast species was used to find sequences that were more likely to be functionally important. Second, a template approach that considered joint positional distributions of word pairs increased the specificity of gene-expression predictions using sequence-based rules. We have demonstrated that higher-order sequence features within TCRs were conserved across multiple *Saccharomyces* genomes. Closely spaced and jointly conserved word pairs were also more likely to be associated with specific gene-expression changes. A large proportion of words contained in templates matched known transcription factor binding sites. In many cases, associations between templates and gene-expression changes were significant in conditions when the corresponding transcription factors are known to be active. In addition, groups of genes that co-conserved both words in a template often were enriched for common functional roles. These results suggest that conserved word-pair templates, which were discovered strictly on the basis of higher-order properties of sequence conservation, also carry biological relevance.

Conserved word-pair templates may be consistent with several underlying biochemical mechanisms. One possible interpretation of templates is that closely spaced sequence pairs may promote direct or indirect interactions between transcription factors by increasing the local concentrations of the individual factors. For example, the proximity of Cbf1p and Met31/32p binding sites may promote interaction between these factors in recruiting their common transcriptional activators, Met4p and Met28p. Experimental studies on the TCRs of *MET3* and *MET28* have demonstrated that the binding of Cbf1p enhances the DNA-binding affinity of Met31/32p [5]. Indeed, biochemical experiments suggest that all these proteins may interact at the TCRs of some sulfur-utilization genes [5].

Another possible regulatory scheme for conserved, closely spaced word pairs is that individual sequences found in templates may correspond to binding sites for transcription factors that bind independently under the same or separate conditions. The Msn2/4p and Hsf1p transcription factors, whose binding sites were similar to words identified in a template, represent an example of multifactorial regulation in response to distinct environmental stimuli [6]. Spacing constraints between their binding sites could nevertheless be important under conditions when both factors are active. Recent experiments have suggested that transcription factors that do not physically interact may still co-activate gene expression as long as their binding sites are spaced within a nucleosome length (approximately 150 bp), due to collaborative competition of the bound transcription factors with core histones [8].

Close spacing between word pairs may be important for reasons other than the promotion of transcription factor interactions. Different regions of TCRs at varying windows away from translation start sites may be more competent at recruiting or inhibiting RNA polymerase. These differences may be influenced by nucleosome accessibility, chromatin structure or DNA physical properties, which can be correlated with local A+T content (see [45] for references). Notably, we have also found that the relative proportions of A and T nucleotides vary considerably within the 200 bp closest to translation start sites (A.M. Moses, M.B.Eisen and Audrey Gasch, unpublished results). Low-complexity words that contained four or more As or Ts could be found in many templates (denoted by T_nC in Figure 4 and Table 1); these words may serve as surrogates for a distance window from the translation start. Binding sites that are close to these low-complexity words may be found in more transcriptionally competent regions of TCRs. Alternatively, the possibility that each word in an identified pair may be found at similar distances from a third conserved sequence element in all TCRs cannot be discounted.

Direct biological models of binding-site organization in TCRs, as exemplified by conserved word-pair templates, provided several advantages over naive statistical models based on sequence combinations in *S. cerevisiae*. Average gene-expression profiles showed that conserved word pairs were more specific predictors of gene expression (giving fewer false positives) than single or pairwise sequences derived from *S. cerevisiae*, indicating that conserved regions among these closely related *Saccharomyces* species were enriched for functional sequences. The consideration of distance constraints between pairs of conserved sequences found many more examples than a previous study of binding-site clustering for multiple transcription factors in *S. cerevisiae* [22]. In addition, we discovered new sequences and pairwise interaction terms using regression models similar to those reported in [20] and [21] (Table 3). Conserved word-pair templates accounted for similar changes in genome-wide expression (*R*-square from approximately 11% to approximately 20%) using only 8 to 10 features, compared with dozens of overlapping features used by other methods [20,21]. Therefore, individual features from our methods were more predictive of genome-wide expression changes.

A key limitation of our approach is the use of hexamers, which may fail to capture known binding sites. For example, the binding sites for Mcm1p and Rap1p are poorly modeled by exact words, in that these transcription factors bind sequences with relaxed specificity at certain positions [11].

Our method missed examples of multifactorial regulation involving Mcm1p or Rap1p that were suggested by previous work using position-weight matrices [17]. In addition, our method required sequence identity for a word to be labeled as conserved. This strict requirement omitted binding sites that may have retained their function, despite mutations in degenerate positions that may have little impact on transcription factor binding. This tradeoff between enumerating all possible words and capturing degenerate positions in binding sites was compounded by the very large number of pairwise word combinations that were enumerated. Further work should incorporate more complicated sequence models, as well as optimization methods that restrict the search space of sequence combinations.

The consideration of joint conservation and close spacing has provided insights into how TCR organization may influence the multifactorial regulation of gene expression in *Saccharomyces cerevisiae*. These criteria were motivated by experimental studies on the positional organization of individual binding sites within TCRs, with the hypothesis that this underlying architecture would be functionally conserved. Even more complicated higher-order sequence rules are apparent in the organization of *cis*-regulatory modules in *Drosophila melanogaster* [46]. Nevertheless, a common organizational theme of the TCRs in both of these organisms is the importance of relative spacing between transcription factor binding sites. The discovery of additional principles for TCR organization will further advance our understanding of how regulatory information is encoded in genome sequences.

Materials and methods

Datasets

Whole-genome shotgun sequencing of *Saccharomyces bayanus*, *S. mikatae*, and *S. paradoxus* has been previously described [26]. All are highly related to *S. cerevisiae*, as they are grouped within the *sensu stricto* branch of the *Saccharomyces* genus [28]. Intergenic regions were aligned using CLUSTALW as described [26] and are available from the Yeast Comparative Genomics website [42]. A total of 4,101 CLUSTALW alignments were analyzed. These alignments were filtered for orthologs in at least three genomes.

Gene-expression measurements were obtained from the Stanford Microarray Database [47] and Rosetta [35]. The main experimental types among the 342 conditions examined include cell cycle [31,32], environmental stress response [30], DNA damage [33,34], cadmium (N. Ogawa and P.O. Brown, unpublished data), and inhibition of ergosterol biosynthesis [35]. This data has been log-transformed (base 2), and each experimental condition has been median normalized.

Dependent conservation of word pairs

To assess whether two words were co-conserved in the same intergenic regions, a chi-square test of independence was systematically conducted for all possible words of length 6. We defined a word to include a 6-bp sequence and its reverse complement. Each transcriptional control region (TCR) for a gene was defined as the 600 bp upstream of its translation start site. TCRs shared between divergently transcribed genes less than 600 bp long were only counted once. A word was labeled conserved in a TCR if all six bases were identical among at least three of the four genomes in the CLUSTALW alignment. For each word pair (*W*, *V*) whose overlap was less than 4, a contingency table $\boldsymbol{C}_{wv}$ was constructed. In this table, $\boldsymbol{C}_{wv}$ = number of TCR($\boldsymbol{I}_{w} \cap \boldsymbol{I}_{v}$), where $\boldsymbol{I}_{w}$, $\boldsymbol{I}_{v}$ are indicator variables for the presence of each conserved word in a TCR, summed over all TCRs. TCRs shared between divergently transcribed genes less than 600 bp long were only counted once. The expected counts $\boldsymbol{E}_{wv}$ were obtained from an independence assumption; that is, the product of the individual word conservation probabilities, multiplied by the total number of TCRs. The chi-square statistic with Yates continuity correction was computed according to the definition:

$$\chi^2_{wv} = \sum_{I_w=0}^{1} \sum_{I_v=0}^{1} \frac{\left(\left| C_{wv} - E_{wv} \right| - \frac{1}{2} \right)^2}{E_{wv}}$$

Spatial proximity of constrained word pairs

The second requirement for a conserved sequence template involved constraints on spatial arrangements between individual words. Any method that evaluates spacing distributions between word pairs must take into account positional biases that may be present for individual words (A.M. Moses, unpublished results). We used a permutation test to evaluate the significance of the median of minimum distances, excluding overlaps, between conserved word pairs. By permuting the TCR labels for one of the words, but not the word positions themselves, we retained the positional biases of individual words within intergenic regions. Within any given TCR $\boldsymbol{t}$, define $\boldsymbol{p}_t(\boldsymbol{W}) = \{\boldsymbol{p}_t^1(\boldsymbol{W}), \ldots, \boldsymbol{p}_t^j(\boldsymbol{W})\}$ as a vector of positions in *S. cerevisiae* where the j occurrences of word $\boldsymbol{W}$ are conserved. Suppose that words $\boldsymbol{W}$ and $\boldsymbol{V}$ were jointly conserved in TCRs $\boldsymbol{T_1 \ldots T_N}$. For each TCR $\boldsymbol{t} \in \{\boldsymbol{T_1 \ldots T_N}\}$, the minimum distance between words $\boldsymbol{W}$ and $\boldsymbol{V}$ was computed as

$$m_t = \min_{j,k} \left| p_t^j(W) - p_t^k(V) \right|.$$

The median of minimum distances, $\overline{D}$, was simply the median of the ordered distribution $\{\boldsymbol{m_1, \ldots, m_t}\}$.

We used a permutation test to generate an empirical null distribution of $\overline{D}$ for all word pairs with $N \geq 10$. After randomly permuting the labels $\boldsymbol{t}$ for the position vectors of word *V*, a permutation test statistic, $\overline{D}^*$, can be calculated as above. By repeating this resampling procedure R times, an empirical null distribution $\overline{D}_{null} = \{\overline{D}^{*1}, \ldots, \overline{D}^{*R}\}$ can be obtained. The significance of the observed median of minimum distances, $\overline{D}$, in

the N promoters was calculated as its quantile in the empirical null distribution $\bar{D}_{null}$. We set an upper bound of $R = 10^6$, but stopped permutations early if 20 or more values in $\bar{D}_{null}$ were found less than $\bar{D}$.

Correction for multiple testing involved control of the proportion of false positives using an FDR method [48]. This method has increased power over Bonferroni-type methods. Permutation quantiles for all N word pairs tested were sorted in non-decreasing order: $\boldsymbol{q_1} \leq \ldots \leq \boldsymbol{q_N}$. Let

$$\boldsymbol{k} = \max\left(\boldsymbol{i} : q_i < \frac{0.05\boldsymbol{i}}{N} \right)$$

Then the first $\boldsymbol{k}$ word pairs in the ordering had a corrected significance level of $q < 0.05$; that is, the rate of false positives is approximately 5%.

Association between template-specified gene groups and gene-expression changes

For each gene-expression condition $\boldsymbol{c}$ in our dataset, $\boldsymbol{c} \in \{1, \ldots, 342\}$, we tested the null hypothesis that a gene subset $\boldsymbol{G}_{wv} \subseteq \boldsymbol{G}$ selected by a conserved word pair (w, v) had the same distribution of gene-expression ratios ($\boldsymbol{E}_{wv}{}^c$) as the entire genome ($\boldsymbol{E}^c$). The alternate hypothesis stated that the two gene-expression distributions were significantly different. Any gene was an element of $\boldsymbol{G}_o$ if its corresponding TCR conserved both sequences in the word pair. Since the size N_o of gene subsets may be small and the distributions may not be normally distributed, we used the nonparametric K-S test. The test statistic $\boldsymbol{K}$ compares the cumulative distribution functions $\boldsymbol{F}_{wv}{}^c$ and $\boldsymbol{F}^c$ corresponding to $\boldsymbol{E}_{wv}{}^c$ and $\boldsymbol{E}^c$ by the formula

$$K = \max_x \left| F_{wv}{}^c(x) - F^c(x) \right|$$

The significance level of an observed value $\boldsymbol{K}^*$ can be obtained using a numerical approximation [49].

A gene subset determined by a word pair was deemed to have significantly different expression if its K-S p-value was less than a certain threshold. To correct for multiple testing, this threshold was established by controlling the FDR. The significance levels $\boldsymbol{p_i}$ from each K-S test were ordered in ascending order. Let $\boldsymbol{N}$ represent the total number of K-S tests performed; that is, the number of jointly conserved, closely spaced word pairs times the number of gene-expression experiments). If $\boldsymbol{k}$ was the largest $\boldsymbol{i}$ such that

$$p_i < \frac{i\alpha}{N}$$

then the first $\boldsymbol{k}$ word pairs in the ordering were deemed to have a significance level of $p < \alpha$.

We ensured that the K-S p-value for the conserved word-pair subset $\boldsymbol{G}_o$ was more significant than subsets $\boldsymbol{G}_w$ or $\boldsymbol{G}_v$ comprised of only one conserved word by computing $\boldsymbol{K}$ for $\boldsymbol{E}_w{}^c$ vs $\boldsymbol{E}_v{}^c$, as well as for $\boldsymbol{E}_w{}^c$ vs $\boldsymbol{E}^c$. The marginal improvement of the joint word pair was defined as: $\boldsymbol{K}(\boldsymbol{F}_o{}^c$ vs $\boldsymbol{F}^c)$ - max($\boldsymbol{K}(\boldsymbol{F}_w{}^c$ vs $\boldsymbol{F}^c)$, $\boldsymbol{K}(\boldsymbol{F}_v{}^c$ vs $\boldsymbol{F}^c)$).

Hierarchical clustering of word pair associations

The $P \times C$ matrix of K-S p-values was log-transformed (base 10), and the word pairs contained in P were clustered by average-linkage hierarchical clustering using the program Cluster [50]. As the log-transformed K-S p-values were all negative, a centered Pearson correlation was used as the similarity metric.

Stepwise linear regression of gene expression

Regression analyses assume that a log-transformed gene-expression measurement, $\boldsymbol{E}_{gc}$ for gene $\boldsymbol{g}$ in condition $\boldsymbol{c}$ can be modeled by a linear equation:

$$E_{gc} = \sum_f M_{fc} \times S_{gf} + \varepsilon_g$$

where $\boldsymbol{S_{gf}}$ represents the score of a sequence feature $\boldsymbol{f}$ in gene $\boldsymbol{g}$, $\boldsymbol{M_{fc}}$ represents the influence term of the feature $\boldsymbol{f}$ on gene expression in condition $\boldsymbol{c}$, and ε_g is the gene-specific error term. Genome-wide expression data was filtered for a set of 4,703 genes with TCRs conserved in three or more *Saccharomyces* genomes. For a certain experimental condition, the list of features was restricted to either two words found in a single word-pair template, or to all words found in conserved word-pair templates that were significantly associated with gene-expression changes in that condition. The score $\boldsymbol{S_{gf}}$ for feature $\boldsymbol{f}$ in a TCR corresponding to gene $\boldsymbol{g}$ was taken as either the number of occurrences in *S. cerevisiae*, or the number of occurrences conserved in three or more *Saccharomyces* genomes. Stepwise linear regression models were fit to genome-wide expression data using the statistical package R [51]. At each iteration, the sequence feature with the largest increase in the R-square goodness-of-fit score was added to the model, where

$$R^2 = \sum_f \left(M_{fc} \times S_{gf} \right)^2$$

Pairwise interaction terms between sequence features $\boldsymbol{f1}$ and $\boldsymbol{f2}$ already selected in the model, expressed as $\boldsymbol{S_{gf1}} \bullet \boldsymbol{S_{gf2}}$, could also be added to the model at each iteration if the features were found in the same conserved word-pair template. Sequence features were added to the regression model as long as the p-values for their associated influence terms ($\boldsymbol{M_{fc}}$) were less than 0.05.

Additional data files

Additional files are available with the online version of this paper. They comprise a tab-delimited text list of 989 identified conserved word-pair templates (Additional data file 1), a figure that shows that incorporation of conservation and word pairs provided more informative average expression profiles (Additional data file 2), a figure that shows that regression models using conserved word pairs represented better fits to genome-wide expression data (Additional data file 3), and the underlying data for Figure 4 that show associations between gene groups that share a conserved word-pair template with gene-expression changes (available as a zip.file, Additional data file 4; all files in TreeView, text tab-delimited, format). When visualized in TreeView (available from [52]), these files correspond to the data underlying Figure 4a and Figure 4b. See Figure 4 legend for details.

Acknowledgements

We are indebted to Audrey Gasch for her insightful advice and suggestions. We thank Justin Fay for a critical reading of the manuscript and help with the permutation test; Peter Bickel and John Storey for their statistical advice; and Nabuo Ogawa and Patrick Brown for sharing gene expression data on cadmium conditions before publication. D.Y.C. is a Howard Hughes Medical Institute Predoctoral Fellow, and M.B.E. is a Pew Scholar in the Biomedical Sciences. This work was carried out under the US Department of Energy contract ED-AC03-76SF00098.

References

1. **Composite regulatory elements: structure, function and classification** [http://www.gene-regulation.com/pub/databases/transcompel/compel.html]
2. Wolberger C: **Multiprotein-DNA complexes in transcriptional regulation.** *Annu Rev Biophys Biomol Struct* 1999, **28:**29-56.
3. Mead J, Bruning AR, Gill MK, Steiner AM, Acton TB, Vershon AK: **Interactions of the Mcm1 MADS box protein with cofactors that regulate mating in yeast.** *Mol Cell Biol* 2002, **22:**4607-4621.
4. Bhoite LT, Allen JM, Garcia E, Thomas LR, Gregory ID, Voth WP, Whelihan K, Rolfes RJ, Stillman DJ: **Mutations in the Pho2 (Bas2) transcription factor that differentially affect activation with its partner proteins Bas1, Pho4, and Swi5.** *J Biol Chem* 2002, **277:**37612-37618.
5. Blaiseau PL, Thomas D: **Multiple transcriptional activation complexes tether the yeast activator Met4 to DNA.** *EMBO J* 1998, **17:**6327-6336.
6. Gasch AP: **The environmental stress response: a common yeast response to diverse environmental stresses.** In *Yeast Stress Responses Volume 1*. Edited by: Hohmann S, Mager WH. Berlin: Springer; 200311-70.
7. Vashee S, Melcher K, Ding WV, Johnston SA, Kodadek T: **Evidence for two modes of cooperative DNA binding in vivo that do not involve direct protein-protein interactions.** *Curr Biol* 1998, **8:**452-458.
8. Miller JA, Widom J: **Collaborative competition mechanism for gene activation in vivo.** *Mol Cell Biol* 2003, **23:**1623-1632.
9. Smith DL, Johnson AD: **A molecular mechanism for combinatorial control in yeast: MCM1 protein sets the spacing and orientation of the homeodomains of an alpha 2 dimer.** *Cell* 1992, **68:**133-142.
10. Drazinic CM, Smerage JB, Lopez MC, Baker HV: **Activation mechanism of the multifunctional transcription factor repressor-activator protein 1 (Rap1p).** *Mol Cell Biol* 1996, **16:**3187-3196.
11. Stormo GD: **DNA binding sites: representation and discovery.** *Bioinformatics* 2000, **16:**16-23.
12. Duret L, Bucher P: **Searching for regulatory elements in human noncoding sequences.** *Curr Opin Struct Biol* 1997, **7:**399-406.
13. Pennacchio LA, Rubin EM: **Genomic strategies to identify mammalian regulatory sequences.** *Nat Rev Genet* 2001, **2:**100-109.
14. Blanchette M, Tompa M: **Discovery of regulatory elements by a computational method for phylogenetic footprinting.** *Genome Res* 2002, **12:**739-748.
15. Loots GG, Ovcharenko I, Pachter L, Dubchak I, Rubin EM: **rVista for comparative sequence-based discovery of functional transcription factor binding sites.** *Genome Res* 2002, **12:**832-839.
16. Schwartz S, Zhang Z, Frazer KA, Smit A, Riemer C, Bouck J, Gibbs R, Hardison R, Miller W: **PipMaker - a web server for aligning two genomic DNA sequences.** *Genome Res* 2000, **10:**577-586.
17. Pilpel Y, Sudarsanam P, Church GM: **Identifying regulatory networks by combinatorial analysis of promoter elements.** *Nat Genet* 2001, **29:**153-159.
18. Bussemaker HJ, Li H, Siggia ED: **Regulatory element detection using correlation with expression.** *Nat Genet* 2001, **27:**167-171.
19. Keles S, van der Laan M, Eisen MB: **Identification of regulatory elements using a feature selection method.** *Bioinformatics* 2002, **18:**1167-1175.
20. Wang W, Cherry JM, Botstein D, Li H: **A systematic approach to reconstructing transcription networks in *Saccharomyces cerevisiae*.** *Proc Natl Acad Sci USA* 2002, **99:**16893-16898.
21. Conlon EM, Liu XS, Lieb JD, Liu JS: **Integrating regulatory motif discovery and genome-wide expression analysis.** *Proc Natl Acad Sci USA* 2003, **100:**3339-3344.
22. Wagner A: **Genes regulated cooperatively by one or more transcription factors and their identification in whole eukaryotic genomes.** *Bioinformatics* 1999, **15:**776-784.
23. Klingenhoff A, Frech K, Quandt K, Werner T: **Functional promoter modules can be detected by formal models independent of overall nucleotide sequence similarity.** *Bioinformatics* 1999, **15:**180-186.
24. Pavlidis P, Furey TS, Liberto M, Haussler D, Grundy WN: **Promoter region-based classification of genes.** *Pac Symp Biocomput* 2001, **6:**151-164.
25. Kel-Margoulis OV, Ivanova TG, Wingender E, Kel AE: **Automatic annotation of genomic regulatory sequences by searching for composite clusters.** *Pac Symp Biocomput* 2002, **7:**187-198.
26. Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES: **Sequencing and comparison of yeast species to identify genes and regulatory elements.** *Nature* 2003, **423:**241-254.
27. Chiang DY, Brown PO, Eisen MB: **Visualizing associations between genome sequences and gene expression data using genome-mean expression profiles.** *Bioinformatics* 2001, **17:**S49-S55.
28. Cliften PF, Hillier LW, Fulton L, Graves T, Miner T, Gish WR, Waterston RH, Johnston M: **Surveying *Saccharomyces* xs genomes to identify functional elements by comparative DNA sequence analysis.** *Genome Res* 2001, **11:**1175-1186.
29. Zhu J, Zhang MQ: **SCPD: a promoter database of the yeast *Saccharomyces cerevisiae*.** *Bioinformatics* 1999, **15:**607-611.
30. Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, Storz G, Botstein D, Brown PO: **Genomic expression programs in the response of yeast cells to environmental changes.** *Mol Biol Cell* 2000, **11:**4241-4257.
31. Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B: **Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization.** *Mol Biol Cell* 1998, **9:**3273-3297.
32. Cho RJ, Campbell MJ, Winzeler EA, Steinmetz L, Conway A, Wodicka L, Wolfsberg TG, Gabrielian AE, Landsman D, Lockhart DJ, *et al.*: **A genome-wide transcriptional analysis of the mitotic cell cycle.** *Mol Cell* 1998, **2:**65-73.
33. Gasch AP, Huang M, Metzner S, Botstein D, Elledge SJ, Brown PO: **Genomic expression responses to DNA-damaging agents and the regulatory role of the yeast ATR homolog Mec1p.** *Mol Biol Cell* 2001, **12:**2987-3003.
34. Lee SE, Pellecioli A, Demeter J, Vaze MP, Gasch AP, Malkova A, Brown PO, Botstein D, Stearns T, Foiani M, *et al.*: In *Biological Responses to DNA Damage Volume 65*. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press; 2000303-314.
35. Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, Armour CD, Bennett HA, Coffey E, Dai HY, He YDD, *et al.*: **Functional discovery via a compendium of expression profiles.** *Cell* 2000, **102:**109-126.
36. Costanzo MC, Crawford ME, Hirschman JE, Kranz JE, Olsen P, Robertson LS, Skrzypek MS, Braun BR, Hopkins KL, Kondu P, *et al.*:

YPD™, PombePD™, and WormPD™: model organism volumes of the BioKnowledge® library, an integrated resource for protein information. *Nucleic Acids Res* 2001, **29:**75-79.
37. Causton HC, Ren B, Koh SS, Harbison CT, Kanin E, Jennings EG, Lee TI, True HL, Lander ES, Young RA: **Remodeling of yeast genome expression in response to environmental changes.** *Mol Biol Cell* 2001, **12:**323-337.
38. Hughes JD, Estep PW, Tavazoie S, Church GM: **Computational identification of cis-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*.** *J Mol Biol* 2000, **296:**1205-1214.
39. Thomas D, Surdin-Kerjan Y: **Metabolism of sulfur amino acids in *Saccharomyces cerevisiae*.** *Microbiol Mol Biol Rev* 1997, **61:**503-532.
40. Fauchon M, Lagniel G, Aude JC, Lombardia L, Soularue P, Petat C, Marguerie G, Esentenac A, Werner M, Labarre J: **Sulfur sparing in the yeast proteome in response to sulfur demand.** *Mol Cell* 2002, **9:**713-723.
41. Williams RM, Primig M, Washburn BK, Winzeler EA, Bellis M, Sarrauste de Menthiere C, Davis RW, Esposito RE: **The Ume6 regulon coordinates metabolic and meiotic gene expression in yeast.** *Proc Natl Acad Sci USA* 2002, **99:**13431-13436.
42. **Whitehead Institute Center for Genome Research - Yeast Comparative Genomics** [http://www-genome.wi.mit.edu/seq/Saccharomyces]
43. Vik A, Rine J: **Upc2p and Ecm22p, dual regulators of sterol biosynthesis in *Saccharomyces cerevisiae*.** *Mol Cell Biol* 2001, **21:**6395-6405.
44. Kwast KE, Burke PV, Poyton RO: **Oxygen sensing and the transcriptional regulation of oxygen-responsive genes in yeast.** *J Exp Biol* 1998, **201:**1177-1195.
45. Liao GC, Rehm EJ, Rubin GM: **Insertion site preferences of the P transposable element in *Drosophila melanogaster*.** *Proc Natl Acad Sci USA* 2000, **97:**3347-3351.
46. Berman BP, Nibu Y, Pfeiffer BD, Tomancek P, Celniker SE, Levine M, Rubin GM, Eisen MB: **Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the *Drosophila* genome.** *Proc Natl Acad Sci USA* 2002, **99:**757-762.
47. Sherlock G, Hernandez-Boussard T, Kasarskis A, Binkley G, Matese JC, Dwight SS, Kaloper M, Weng S, Jin H, Ball CA, *et al.*: **The Stanford Microarray Database.** *Nucleic Acids Res* 2001, **29:**152-155.
48. Benjamini Y, Hochberg Y: **Controlling the false discovery rate: a practical and powerful approach to multiple testing.** *J R Statist Soc B* 1995, **57:**289-300.
49. Press WH, Teukolsky SA, Vertterling WT, Flannery BP: *Numerical Recipes in C* 2nd edition. Cambridge: Cambridge University Press; 1992.
50. Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and display of genome-wide expression patterns.** *Proc Natl Acad Sci USA* 1998, **95:**14863-14868.
51. **The R Project for Statistical Computing** [http://www.r-project.org]
52. **Eisen Lab** [http://rana.lbl.gov]